2012-12-10

Predicting Stock Market Using Online Communities Raw Web Traffic Streams

Pierpaolo Dondio

Technological University Dublin, pierpaolo.dondio@tudublin.ie

Recommended Citation

Dondio, Pierpaolo, "Predicting Stock Market Using Online Communities Raw Web Traffic," (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on , vol.1, no., pp.230,237, 4-7 Dec. 2012 doi:10.1109/WI-IAT.2012.206

Predicting Stock Market Using Online Communities Raw Web Traffic Streams

Pierpaolo Dondio
School of Computing
Dublin Institute of Technology, Kevin Street, Dublin, Ireland
Pierpaolo.Dondio@comp.dit.ie

***Abstract*— This paper investigates the predictive power of online communities traffic in regard to stock prices. Using the largest dataset to date, spanning 8 years and almost the complete set of SP500 stocks, we analyze the predictive power of raw unstructured traffic by filtering stock daily returns with traffic features. Our results partially challenge the assumption that raw traffic simply trails stock prices, as expected from a noisy signal without the sentiment direction. Raw traffic is shown to predict prices with statistical significance but with small economic impact. Anyway, this impact rises to moderate under the following conditions: 3 to 7 days lag and stable traffic level. Moreover, the quality of the predictions significantly increases when a high level of traffic is coupled to low market volatility, while a high level of traffic in period of high volatility usually denotes late reactions to violent market movements and a consequent poor predictive power. The findings set interesting future works in the definition of novel indicators for market analysis based on web traffic analysis, to be coupled with complementary tools such as sentiment analysis.**

Keywords: Online communities, Stock Market, Predictive model

I. INTRODUCTION

Since their inception, online communities about finance have received a growing attention as a valid source of market analysis, and they have gradually gained credibility. Despite this clear trend, evidence regarding the predictive value of financial social media is not definitive. In one of the earliest papers by Antwellier [2], the author concludes how the impact of the message board is significant statistically but not economically, while more recent results report accuracy in the range of 70-80%. Moreover, all the previous studies do not cover a period of evaluation of more than 1 year and - except in one case - no more than 45 stocks. This paper contributes to the debate about whether online communities have predictive market ability. We propose an evaluation using the most extensive dataset to date, in terms of time span - 8 years - and stocks number - about 480.

We identified 3 major techniques and 3 levels of analyzing social media content for market predictions.

The first source is the unstructured stream of web-traffic produced by the community. In its essential model, it is a stream of messages (posts, twits) tagged with three dimensions: user, time, stock associated.

The second source of information is represented by text-based features, typically an indicator of the sentiment expressed. The previous literature is dominated by such approach. Market prediction models are based on a sentiment index that gives the daily raw traffic a positive/negative direction. Nevertheless, text features are not limited to sentiment. Bollen [2] experiments with 7 text-based features, encompassing things such as calm.

Third, other features come from behavioural/social information rather than text, such as the reputation of the individual in the community, profile, friends, the way he interacts with other members.

Given these 3 sources of features, it is possible to aggregate them at user-level - where each user is considered to have a different impact on the overall index –; at community-level (where predictions are generated by considering all users the same) and at multi-community level. This study concerns the investigation of web traffic quantitative data at community level, a complementary research to usual text-based analysis. We first pose the following research question:

Can patterns of raw traffic predict market? Under which conditions?

The answer to the above question seems an obvious no. Unqualified traffic is too noisy and, more importantly, it has no direction in terms of the positive/negative sentiment. How can we predict something we do not understand?

There are interesting considerations that justify the question as a valid research question (see section 2), mainly on the ground that under some conditions traffic could act as a proxy, an approximation or even a substitute for users' sentiment and from the fact that how users generate their own traffic is nothing but random.

The contribution of our paper is the effort to produce an answer to the above question. Rather than provide definitive evidence, the paper provides enough encouraging evidence to justify further investigations into the definition of novel indicators complementing existing text-based ones.

We also contribute with the largest dataset, filling a gap in previous experimentations where either the time span or the stock set was extremely small.

The paper is organized as follows: in section 2 we discuss why the hypothesis of raw traffic could be reasonable, in section 3 we describe our methodology of analysis; evaluated in section 4, section 5 describes a further experiment using a longer time range while section 6 presents related works to date.

II. A CASE FOR RAW TRAFFIC

In this section we discuss few reasons why it is worthy to investigate the predictive power of raw unstructured traffic. The main idea is that traffic could act as a proxy, an approximation or even a substitute for users' sentiment. Recent works seem to back the validity of the hypothesis. We list our considerations.

A. *Direct evidence collected via surveys*

We conducted a survey on the website FinanzaOnline.it [6], the largest Italian online community with about 120 000 registered users and about 15 million post. We asked the following questions:

Q1: If you write on a stock board, do you hold the stock? If not, why are you writing there?

Q2: Do you still write about stocks you have sold?

We collected about 350 answers. The results show how 78.7% of users replied *yes* to the first question, adding as most frequent comment that, if they are writing on a stock they do not hold, the majority of time it is because they are considering buying it. Users also replied how the activity fades after the stock is sold. It is reasonable to presume that users' activity is nothing but random. The results of the surveys allow us to believe that traffic could act as an indicator of community interests towards specific stocks, and therefore has some kind of predicting capability. The key question is therefore the following: *is this kind of interest/association between traffic levels and stocks enough to make market predictions*?

B. *Absence of sentiment*

Another reason to consider raw traffic data is that the large majority of messages are out-of-topic, containing no sentiment at all. Anyway, it is a reasonable hypothesis that the presence of such messages about a specific stock at a specific time and market condition is not random. It is also common that users never publicly express their sentiment.

C. *Positive biased and technical reasons*

There is evidence over a strong positively-biased sentiment populating financial on-line communities (see [10]), that allows us to presume that traffic could be a proxy for at least positive sentiment. Messages on average are strongly over-bullish. This suggests that the predictive value of web-traffic, if any, could result asymmetric, i.e. effective in one direction only, either buy or sell.

Partially, the three above observations find a confirmation in the work by Bollen [3]. Bollen reports that it is not the positive/negative sentiment that predicts the market, but actually one particular mood extracted by the text that he calls "*calm*". A reasonable hypothesis is that calm is a concept that can be also effectively identified by patterns of traffic as well. The work by [5] provides further evidence about making good prediction without sentiment. Using a limited dataset of 4 stocks, the author concludes how market movements can be predicted with an 80% accuracy by relying on non-textual blogs dynamics such as increase in blog comments, average response time, quotations, length of comments.

III. PREDICTING WITH RAW TRAFFIC

Our dataset is composed by a stream *M* of meta-data about messages posted on Yahoo! Finance. *M* is a sequence of tuples (u, s, t) associated to each message, where $u \in U$ is the user author of the message, $s \in S$ is the stock the message refers to, t is the time of message creation. We collected about 26 millions tuples from Yahoo! Finance, spanning 8 years and 478 out of 500 stocks of the US SP500 index. The stream *M* identifies a 3-dimensional space with dimensions stocks (S), users (U) and time (T). The time dimension T is discretized by choosing an interval of time ΔT. In our simulation ΔT is always equal to one day, meaning that we do not study intraday trading.

Distinct to the stream *M* is a function $P(s,t): S \times T \rightarrow \mathbb{R}^+$ that associates the stock closing price to each stock and day. We use the closing price adjusted for dividends and share splits, using Bloomberg as a source.

By partitioning the stream *M* we can isolate data regarding a single stock or user in a particular interval of time. For the remaining of this work we need to define the following time series:

$N_{u,s}(t)$ = n. of messages of user u on stock s at day t

$N_s(t)$ = n. of messages by all users on stock s at day t

$N_u(t)$ = n. of messages by user u at day t (on any stock)

A. *Predicting the stock price*

Our scope is to study the correlation between $N_s(t)$ and stock *s* prices. We seek to verify if web traffic time series – or some subset of it – can predict price movement.

Anyway, a straightforward correlation between a traffic signal such as $N_s(t)$ and stock *s*' historical prices does not achieve any clear result and it is hard to be meaningfully analyzed (see also [2]). Over our dataset, a direct correlation between $N_s(t)$ and stock *s* historical prices has an overall negligible value of 0.038. A correlation between the entire web-traffic signal and stock prices has little hope. Web traffic has different dynamics than market prices: it is sparse, it has long periods of little or no signal interrupted by high peaks, while stock prices are more continuous, trend- based integral signals.

Figure 1 shows an example of the stock Boston Property (BEXP), an SP500 property firm of about 16 B$ capitalization, while figure 2 shows its price chart.

We also stress that a viable trading strategy seeks precision rather than recall. It is not required providing a prediction at every interval, but rather provide precise recommendations when certain conditions are met.

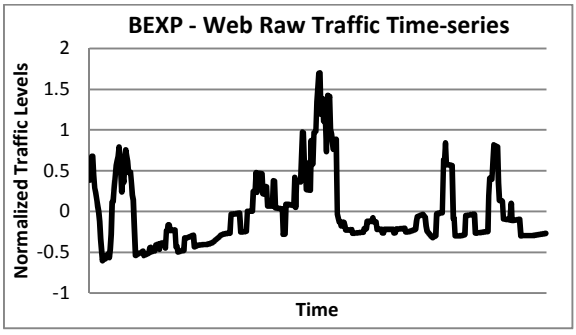


Figure 1. Traffic Chart for BEXP

We believe that what needs to be correlated with market prices are some *features* of $N_s(t)$, and testing if the presence

of such features are an indicator of higher or lower than average returns.

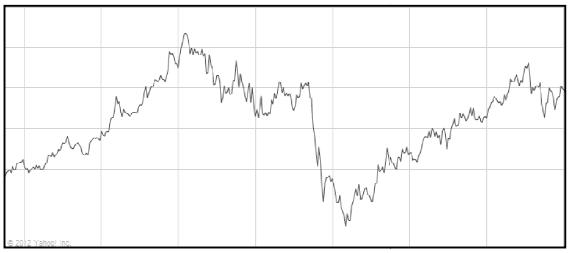


Figure 2. Price Chart for BEXP

We believe that what needs to be correlated with market prices are some *features* of $N_s(t)$, and testing if the presence of such features are an indicator of higher or lower than average returns.

We start by defining $Z_{s,d}$, that is N_s normalized with a standard score obtained using an average and a standard deviation computed over a time-window of d days before. We call d the memory size. Therefore:

$$Z_{s,d}(t) = \frac{N_s(t) - \mu_{N_s(t-d,t)}}{\sigma_{N_s(t-d,t)}} \qquad (1)$$

Using the function $P(s,t)$ we also define:

$$G_s(t) = \frac{P(s,t+1) - P(s,t)}{P(s,t)} \qquad (2)$$

that represents the daily return for stock s at day t. In order to isolate traffic features of interests, from $Z_{s,d}$ we define a function called *signal function* S, that is a binary time-series equal to 1 when the traffic function $Z_{s,d}$ satisfies certain criteria. For instance, if the feature we are interested is '*traffic value is above a certain level'*, we set a threshold over the values of $Z_{s,d}$, defining the following binary signal function:

$$S_T = \begin{cases} 0 \text{ if } Z_{s,d} & < T \\ 1 \text{ if } Z_{s,d} & \geq T \end{cases}$$

Therefore S_T filters $Z_{s,d}$ and considers only days with certain level of traffic. We now apply a cross-correlation operator to S_T and G_s following a methodology similar to (Gruhl et al. 2005). We call Tr the cross-correlation-like coefficient between the time series G_s and S_T:

$$Tr(n,s,d) = (S_T * G_s)(n) = \sum_{t=0}^{\infty} S_T(t)G_s(t-n), n < 0$$

$$Tr(n,s,d) = (S_T * G_s)(n) = \sum_{t=0}^{\infty} S_T(t+n)G_s(t), n \geq 0$$

Note how we made explicit the fact that Tr is a function of n (the lag), stock s and memory d.

What is the meaning of Tr?

As for each cross-correlation coefficient, a high value for negative lags n means that the first series (traffic series S_T) leads - anticipates - the second (price returns) and viceversa for negative lags - the first series is trailing the second. By using S_T instead of $Z_{s,d}$ we gave a direct trading-related meaning to the cross-correlation coefficient. Since S_T is a binary series, the coefficient Tr is composed by zeros (when $S_T(t) = 0$) or price returns $G_s(t)$ (when $S_T(t) = 1$). Therefore Tr is a sum of daily returns.

The final value is equal to the total returns (gain) of a trading strategy for stock s which:

1) buys s on the closing bid for each day where $S_T = 1$

2) invests the same amount of capital (and therefore we buy a variable number of shares)

3) sells the stock at the closing bid of the following day

We note how this is not an efficient trading model, but rather a way to give a more understandable economic meaning to the number obtained. Each active day (where $S_T = 1$) a fixed commission fee would be paid. When there are 2 consecutive active days, the strategy acts like the stock is sold and bought at the same closing bid, which represents a net loss in commissions. A more correct approximation would therefore neglect commissions every time there are two consecutive days where $S_T = 1$. Anyway, it is not the scope of this paper to optimize trading strategies, but rather test if community traffic has predictive power compared to a market benchmark that we define in the following section.

B. Setting a Market Benchmark

The product $S_T G_s$ over all the time periods identifies a sample p_T of certain size m selected from the underlying price return time series G_s. In other words S_T selects a set of m daily returns form G_S.

Since the distribution G_S represents the market price, and by using S_T we selected a sample p_T from G_S, we wonder if this sample is better, worst or statistically equal to any sample we could draw from G_S, i.e. from any random sample of market prices.

Our way to build a market benchmark is therefore to test if the sample p_T significantly deviates from the sample distribution of same size (let's refer to the size as m) obtained from G_s, a distribution with average μ_{G_s} and standard deviation $\sigma_{G_s}/\sqrt{m}$, where μ_{G_s} and σ_{G_s} are the average and standard deviation of G_s.

We therefore performs a two-tails test checking the null hypothesis $\mu_{Z_T G_s} = \mu_{G_s}$. Note how we rely on the fact that G_s, being the daily rate of return of a stock, is normally distributed. We summarized our methodology:

1) We compute G_s for stock s - the daily return (normally distributed)

2) We compute $Z_{s,d}$

3) We set some features conditions on $Z_{s,d}$ and we compute the binary time series S_T

4) We cross-correlate $G_s$and S_T

5) We collect all the non-null terms of the cross correlation. Let's presume the terms form a sample set we call set p_T of size m.

6) We perform a two-tails test between the sample distribution of size m of G_S and the sample p_T

We classify the outcome of the statistical test by comparing the two means and considering the test *p-value*. First, if the average of the sample p_T is above the mean μ_{G_s}, we classify the test as positive, otherwise as negative. Positive therefore means that the raw traffic sample outperformed the market and viceversa for negative tests. We then consider the *p-value* of the test. If the *p-value* is above 0.9, we further classify tests in abnormally positive or abnormally negative. The meaning of abnormally positive is that in the test the traffic outperformed the market with high statistical confidence and viceversa for abnormal negative results.

IV. Experimental Analysis of Tr coefficient

We performed simulations on all the stocks available using the following parameters to normalize the traffic series and generate the signal function S_T: memory d to generate $Z_{d,s}$ in the set $\{20, 30, 60, 90, 120, 240\}$ days, and threshold from 0 to 3 units with step 0.1 We tested with lags from -10 to 10 days. If not stated differently, results are the average of all the simulations.

We computed our cross-correlation coefficient Tr and we tested the statistical difference between the traffic-based average values (referred as the *traffic*) and the sample-distribution of returns G_s, referred as the *market* (benchmark) computed as described earlier. We tested the following market features: traffic above a threshold, between a threshold interval, traffic increments, decrements and absolute variations, as explained later.

The first graph 1 shows, by lag, the percentage of time traffic outperformed market. We notice:

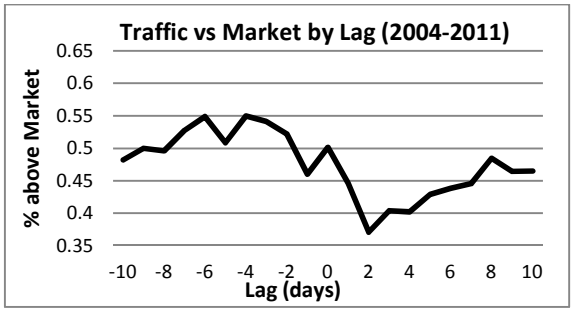
1. for positive lags, traffic constantly and considerably underperforms market. Negative falls of prices seem the driver of high traffic.
2. for negative lags, traffic is above the market for most times. The best performance is obtained with a lag of 7 to 3 trading days before the price series, with an average of 54.5% of simulations above the market.

The delay of 7 to 3 days is in accordance with what reported by Spiegel [10] and Gu [8] where the maximum return was obtained from 5 to 3 days before the price event.

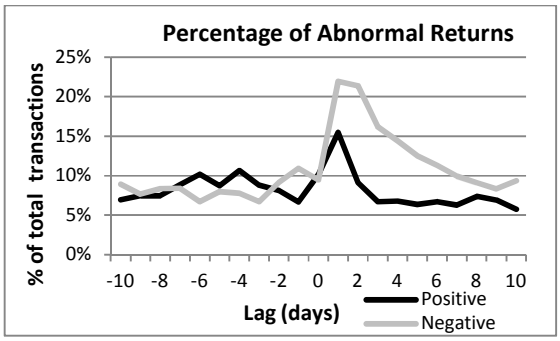
Graph 2 shows the distribution, by lag, of positive and negative abnormal returns as a percentage of the total number of simulations.

Graph 2 clearly shows how abnormal returns (positive and negative) are concentrated on the day (lag=0) or the day following a big price movement. The large amount of abnormal negative returns could be the effect in the dataset of the credit crunch fall.

Graph 1 showed that, within 7 to 3 days lag, on average traffic beats the market 54.5% of the time. The next graph 3 helps understanding if traffic returns are economically significant. Graph 3 shows the performance of traffic vs market after commissions, represented by the new grey dotted market line that is now *harder* to beat. If without commissions traffic beats market in 9 out of 11 negative lags, it now outperforms the market only in 4 lags in the region -6, -3. In that region traffic still outperforms the market but now on average 52.2% of the time against 54.5% without commissions. The gap is still statistically significant, even if the economic impact is sensibly reduced. How did we estimate the impact of commissions?

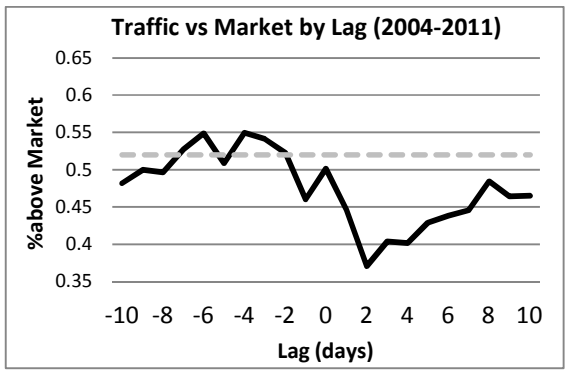


Graph 1



Graph 2

If we consider a 7$ fixed commissions (offered by Etrade, ScottTrade), a capital of 10,000$ each transaction, and since from our data the average length of consecutive trading days (where $S_T = 1$) is equal to 8.25 days, in a real trading implementation there would be a buy and sell every 8.25 days, and the daily commissions cost on each transactions can be estimated as: $\frac{2*0.07}{8.25} = 0.017\$$ for each dollar spent, a percentage that closes the gap and make the market line closer. If we increase the commissions, the gap between *traffic* and *market* is narrow.

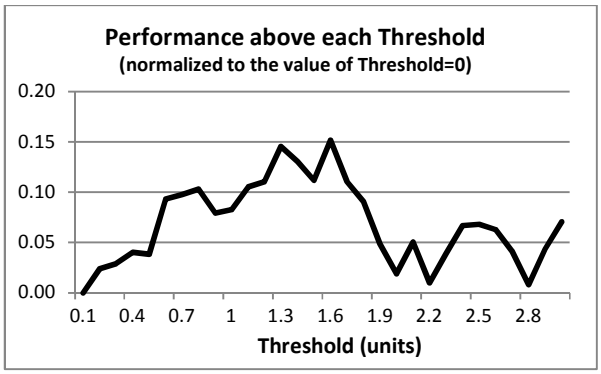


Graph 3

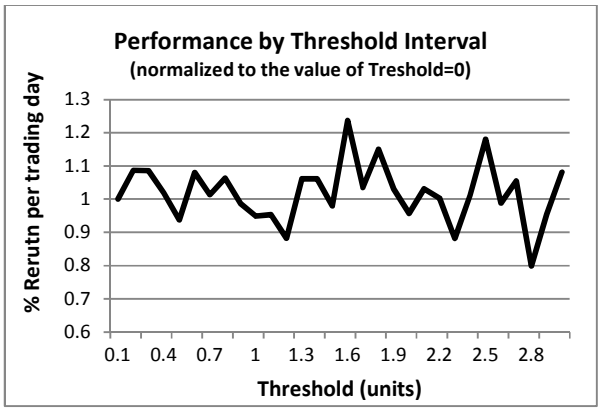
Our first conclusion is that our correlation coefficient Tr has showed how the *raw traffic is in general market efficient*. Anyway, there is a region where it statistically outperforms the market with small economic return, and leaves the room open for the exploration of specific features that might reveal bigger inefficiency.

A. Threshold Analysis

In the following analysis, as a measure of performance we use the ratio of times that *traffic* signals beats the *market* benchmark. The following graph 4 shows the performance of the feature *traffic above a certain threshold*, where threshold varies from 0 to 3 at intervals of size 0.1. Graph 4 values are normalized over the value of threshold equals to 0. The graph shows an interesting sensitivity to the threshold with a steady increase when the threshold increases, up to a 15% gain around a value of 1.3-1.6; then a steady sharp decline followed by an unstable behavior. An interpretation could be that high traffic levels predict the market better than just-above the average levels, but when traffic levels are abnormally high they usually result in poor performance and they represent late abnormal reactions to big price movements.



Graph 4



Graph 5

We now study the feature *traffic in a threshold interval*, i.e. we wonder the behaviour of the raw traffic coefficient not above a threshold but in an interval $[T_a, T_b]$. We modify S_T as follows:

$$S_{T_{intervals}} = \begin{cases} 1 & \text{if } T_a < Z_{s,d}(t) < T_b \\ 0 & \text{elsewhere} \end{cases} \quad (4)$$

Graph 5 shows the return in each threshold interval of size 0.1 normalized with the value of the first interval $[0,0.1]$. Best performances are achieved with a threshold around 1.4 to 1.6. Very high threshold values produce variable returns such as the very low return at 2.8 and very high at 2.5.

B. Analysis of Deltas

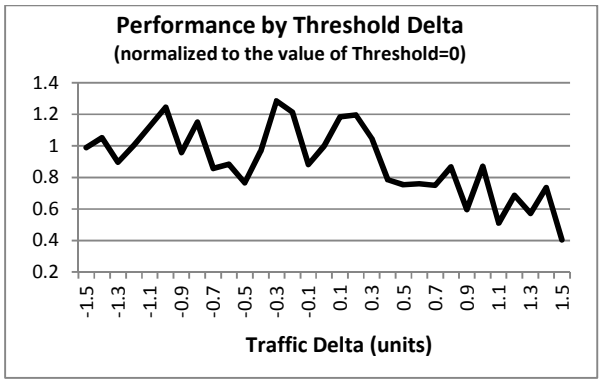
It is interesting to study not only the application of a threshold over absolute values of traffic, but also a threshold on the variations of traffic level. The feature is defined as *traffic daily increment/decrement (deltas) in a threshold interval*. In order to catch this feature the binary time series S_T is therefore redefined as follows:

$$S_{T_{delta}} = \begin{cases} 1 & \text{if } T_a < Z_{s,d}(t) - Z_{s,d}(t-1) < T_b \\ 0 & \text{elsewhere} \end{cases} \quad (5)$$

The following graph 6 shows returns per delta level, i.e. at each value x the graph shows the return when the traffic level varied by a delta included in $[x, x + 0.1]$. The graph shows how negative deltas – traffic decreasing – has more predictive value than positive deltas. If we aggregate the values we obtain that negative deltas outperform positive ones by 29.6%. It is interesting to notice how sudden increases in traffic perform poorly - i.e. they usually predict a price fall - and best performances are centered around 0, where the traffic is stable.

Finally, we can also consider absolute values of delta, meaning that we allow the series to change in both directions:

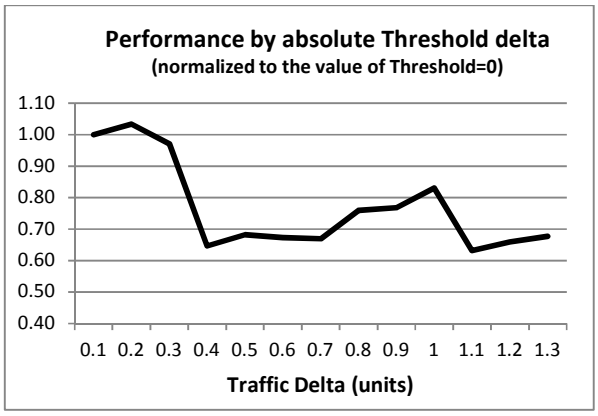
$$S_{T_{abs}} = \begin{cases} 1 & \text{if } T_a < |Z_{s,d}(t) - Z_{s,d}(t-1)| < T_b \\ 0 & \text{elsewhere} \end{cases} \quad (6)$$



Graph 6

Graph 7 shows how performances are significantly better when the traffic varies less than 0.3 units (now in both directions), while performances decline sensibly for larger variations. We regard this as another interesting conclusion: *stable levels of traffic rather than sudden changes predict the market better* than rapidly increasing traffic, as a naïve hypothesis would have suggested. This interesting result is

in accordance with Boella's key claim that *calm* is the mood of online users that better predicts market trends.



Graph 7

C. 4.3 Historical analysis

Our dataset allows us to analyze the behaviour of traffic over almost 8 years. Moreover, these 8 years include a period of stable bullish market (up to 2007), the credit crunch (2008-half 2009), a rapid rally (2009-2010) with high volatility. Graph 8 shows the return by lag for each year. Table 1 shows, for each lag, the number of years – out of eight – where traffic overall beat the market benchmark, while table 2 shows, for each year, the number of negative lags – out of 11 – where traffic beat the market.

TABLE I. YEARS ABOVE THE MARKET

LAG	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
Years above the market	5	5	5	7	6	5	6	5	4	5	4

TABLE II. LAGS ABOVE THE MARKET

Year	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**
Negative lags above market (out of 11)	7	11	3	9	8	0	11	11

It is interesting to notice that the traffic failed to beat the market in 2009 and partially in 2006. 2009 was a year of violent and fast recovery that the traffic failed to capture. On average, returns for negative lags were below the market only in 2009, flat in 2006 and higher in the remaining 6 years. Both 2006 and 2009 were years of easy trading with stable bullish conditions. None of the negative lags outperformed the market in 2009, while all of the 11 outperformed it in 2005, 2010 and 2011. Good results are constantly obtained in a region between 7 and 3 days lag.

In conclusion, we had shown how community traffic is overall *market efficient*, but some of its features can predict the market with statistical significance and moderate economic impact for a large subset of traffic where:

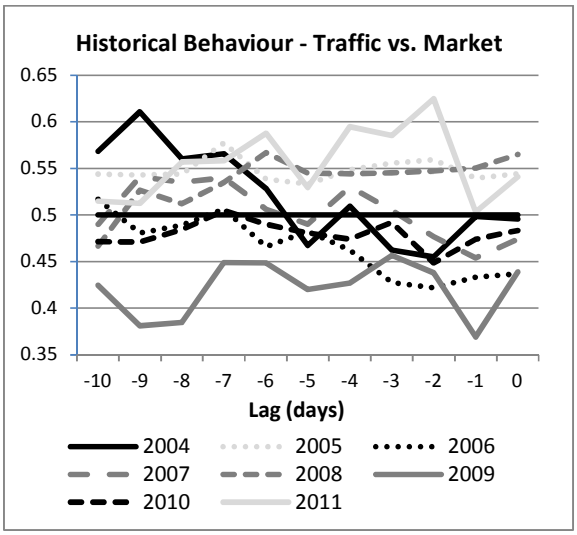
1) the lag goes from 7 to 3 days

2) the traffic does not vary too much

3) the traffic level is above the average but not abnormally

Regarding historical returns, the traffic performed poorly in condition of high volatility and bullish market as 2009, but it shows significant good performance in other periods, including the strongly bearish 2008 market. The hypothesis of market efficiency has been confirmed for the whole traffic, but there exists a subset of traffic patterns where there is statistical deviation from market efficiency and the economic return could justify a trading strategy. For instance, a trading strategy could exploit the inefficiency between different threshold levels, or between different lags or buy in correspondence of medium and stable traffic.

It is also interesting that our conclusions are similar or improved as compared to findings in literature (see related works section) that are usually ascribed to sentiment-based indexes rather than raw traffic.



Graph 8

V. MID-TERM PREDICTIONS WITH PRICE INFORMATION

In this section we report a further experimentation, where we try to predict if the stock price will rise or fall by a fixed percentage. This equates to set a fixed symmetric *stock profit* and *take loss* in a trading strategy. We perform the test to check the predictive ability in a mid-term temporal range; it could be the case that daily returns are too hard to predict, but a mid-term trend could be more easily spotted. Moreover, the test is definitely closer to a real trading scenario and it allows us to better test trading potential.

We proceeded as follows. We first considered a raw-traffic indicator for each stock that takes in consideration both the level of historical traffic on a stock s and the absolute level of traffic of s *compared* to all the other stocks that day. Given a stock $\bar{s}$ and a day $\bar{t}$, we considered the usual $Z_{\bar{s}}(\bar{t})$, the *z-score* of the historical time series $N_s(t)$ computed in $\bar{t}$, and this time we also considered $Z_{\bar{t}}(\bar{s})$, that represents the *z-score* of the distribution of $N_s(t)$ varying stocks instead of the time dimension t(that remains fixed).

Therefore $Z_s(t)$ expresses the level of traffic in respect to stock s history (the distribution goes across the *time* dimension), while $Z_t(s)$ expresses the level of traffic of stock s in respect to all the other stocks that day (the distribution goes across the *stock* dimension). The indicator $I(s,t)$ is defined as the geometric mean of the two *z-scores*:

$$\mathrm{I(s,t)} = \mathrm{Z_s(t)Z_t(s)} \qquad (7)$$

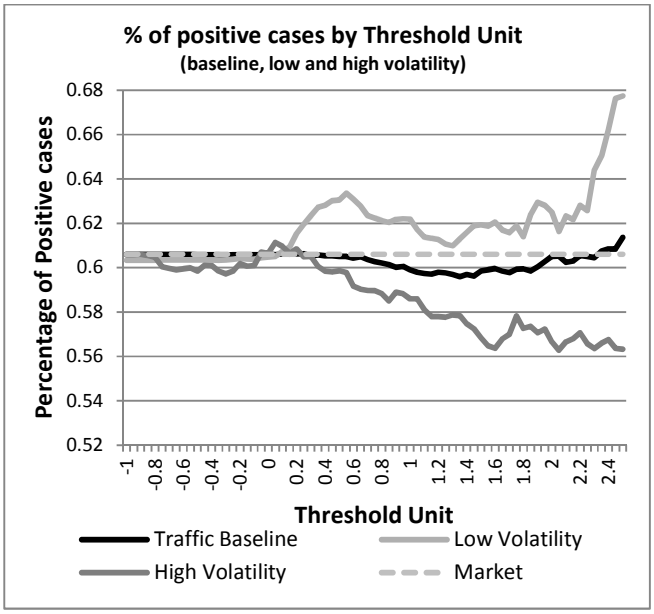
For each stock s we marked each trading day t as *positive* or *negative* according to which of the following two events happened first: (1) the stock price rises more than a fixed percentage g or (2) the stock prices falls further than g. We then tested whether our web-traffic indicator I was able to predict positive or negative movements. We performed experiments with a 10% fixed symmetric target price.

Following the same methodology we used for the signal function S, we studied the behavior of $I(s,t)$ above a certain threshold, reporting for each threshold the percentage of positive outcomes. We also tested if this percentage statistically differed from the *market benchmark* percentage, obtained using all the stocks and trading days.

The black line of Graph 9 shows results obtained by varying the threshold in [-1 , 2.5] by intervals of 0.05 unit (defining a total of 70 bins) and by considering all the stocks and trading days. Results are expressed as percentage of positive cases. The flat dashed grey line represents the market benchmark, equals to a percentage of 0.606. We are looking for threshold intervals where the black line (*traffic*) significantly diverges from the market line. Graph 9 shows how performances do not vary sensibly from the market benchmark. Few statistically significant results (3 out of 70 total threshold settings) are obtained in the region around 1.3, where the *traffic* line is below the market line, meaning that the *traffic* is effective in predicting negative outcomes (price falls). Anyway, the rate of such predictions does not provide the basis for a viable trading strategy after commissions are deducted (we note that, since we predict a price fall, the commissions would be more expensive).

The situation changes when we take in consideration price-related information. Our hypothesis is that the predictive value of *traffic* might be increased considering price information as well. If the stock price is raising or plumbing rapidly (usually index of high volatility), then the traffic high levels might just be reactions to these violent price movements and, as seen in graph 1 and 2, generate late reactions and insignificant returns. This would confirm the analysis of Antwellier [2] as well. Anyway, exceptional high levels of traffic during periods where the stock price is mainly flat and the volatility is low are worthy to be investigated. The dark and light grey lines of Graph 9 and table 3 show the results of two experimental settings that implement the above idea. The first (light grey line) considers only trading days of those stocks whose price varied in the interval [-3%,3%] in the last week *and* month (i.e. the price was relatively flat), while the second (dark grey line) considers the dual set of trading days, where the stock prices varied of more than 3% in both directions.

If we consider the light grey line of Graph 9 (situation of low volatility), we see a clear positive trend after the level of traffic rises above 1.5 unit. Table 3 reports now 17 threshold values where traffic outperforms positively the market benchmark, mainly in the region of high traffic values. The dark grey line exhibits a dual trend, with an increasing ability of predicting price falls when traffic levels are high. The analysis shows how, in situations of low volatility, a high level of web traffic for a stock means a *buy* signal while it acts as a *sell* signal for high volatile markets.



Graph 9

TABLE III. ABILITY TO PREDICT INCREASE ON PRICE (COUNT OF NUMBER OF BINS OF 0.05 UNIT SIZE DEFINED OVER THRESHOLD VALUES)

	Baseline	**Low volatility**	**High volatility**
Better than market	0	17	0
No significant difference	67	53	57
Worse than market	3	0	13

VI. RELATED WORKS

This paper investigates the predictive power of online communities data with respect to financial trading.

The issue has been first extensively by Antweiler and Frank in [2]. The dataset used was 1.5 million posts from Yahoo Finance and RagBull, and the study covered 45 stocks of the Dow Jones. The authors applied text-mining techniques - a naive Bayes classifier - to extract a polarity sentiment from users' posts. The authors' key conclusion was the following: the effect of stock messages helps predict market volatility, but the effect on stock returns is statistically significant but economically moderate.

Spiegel et Al. [9] investigated the effect of rumours over stock returns. In their context, rumours are not coming from online communities and they are not user-generated, but rather news, recommendation and indications coming from financial portal such as *The Bursa* (www.dbursa.com) or *Trading for Living* (trading4living.com). The study concludes how during the event day and the 5 days

preceding it the abnormal stock return is positively and statically significant. The dataset was composed by 958 Israeli stocks monitored for 27 months using a set of about 2000 rumours.

The recent work by Bollen [3] investigates the predictive power of Twitter's messages. The dataset used consisted of about 10m posts by 2.7M users in the period February-December 2008. The trained system was then tested over one-month-period in December 2008 over the closing of the Dow Jones Industrial index. The methodology used was as follows: from tweets' texts authors extracted 7 indicators of mood using OpinionFinder and GPMOS. Using a Granger causality analysis, authors correlated DJIA values to GPOMs and OF values of the past n days to obtain 83% accuracy. The author reports that calm, other then positve/negatie sentiment better predicts the market.

The work by De Choudhury et al. [6] is of particular interests, since it derives market predictions by analyzing communities' dynamics rather than text. The authors focus on blogs and they identify a set of dynamic features, such as normalized response time, early and late responses, and activity measurement such as activity loyalist and outliers. Other features are post length, rank - as provided by the blog editor software, number of posts and comments. These features are then correlated to the market dynamics training a support vector machine with the following results: 78% accuracy in predicting the magnitude of the movement and 87% for weekly movement.

Similar works in the area are the ones by Agarwall et al. [3] on the general problem of identifying influential bloggers in a community and the work by U. Zhang [5], that studied the correlation between past-performance of a user and its reputation. The authors provide insight on what constitutes a reputable and respected user, and conclude how reputation derives from a more complex synthesis of various behavioural factors besides its textual contributions, implicitly confirming the validity of non-textual features.

In conclusion, the panorama is dominated by text-mining technique and past-performance indicators based again on sentiment explicitly tagged. There is a mixed set of conclusions about the predictive capacity of online communities, ranging from not economically significant impact to highly significant impact. All the studies, except one, covers 1-year period or less, and no more than 45 stocks and only [5] provides behavioural elements that are then correlated to the stock market.

VII. Conclusions And Future Directions

The paper has investigated the predictive power of online community traffic. We believe to have provided enough encouraging evidence to justify further investigation. We have identified three main future works area:

1) Data Mining techniques. This first paper has proposed a simple filter based on web-traffic features applied over the daily return time-series. In order to unveil complex interactions between market indicators such as prices, company fundamentals and traffic features, an ongoing study is applying classifiers and clustering techniques to our dataset

2) User-level analysis. The idea is to investigate the following research question: *are there users whose patterns of traffic constantly outperform/underperform their peers?* We wonder if there are users whose patterns of traffic help to increase the predictive capacity. The hypothesis of the existence of such sets is valid. Market efficiency might still be valid for the whole community of traders, but not in specific subsets of it.

3) Behavioural indicators. Future works should be directed towards the definition of market indicators considering also behavioural features of users in the community rather than solely traffic patterns.

Regarding our conclusions, in this study we have shown how raw traffic predicts the market with statistical significance but with average small economic impact after commissions. Anyway, the economic impact is moderate for a large subset of traffic identified by the following conditions: 3 to 7 day lag, stable traffic with high but not abnormal values. In the second part of our analysis we have shown how there is a subset of users that constantly outperforms the others. The findings are in line or outperform the ones reported in literature using sentiment-analysis-based algorithms, and we believe to have provided enough evidence to set the foundation of future works in the development of novel market indicators. Much work has to be done in the definition of implicit behavioural models able to approximate online users' intentions.

References

[1] Agarwal N., Huan Liu, Lei Tang, and Philip S. Yu. 2008. Identifying the influential bloggers in a community. In Proceedings of the international conference on Web search and web data mining (WSDM '08). ACM, New York, NY, USA, 207-218.

[2] Antweiler, W. and M.Z. Frank, 2004. Is all that talk just noise? The information content of internet stock message boards. Journal of Finance, 59: 1259-1295.

[3] Bollen J., ,Twitter mood predicts the stock market

[4] Cook, D.O. & Lu, X. 2009 Noise, Information, and Rumors: Internet Board Messages Affect Stock Returns. Working Paper.

[5] De Choudhury Munmun, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. 2008. Can blog communication dynamics be correlated with stock market activity?. In Proceedings of the 19th ACM conference on Hypertext and hypermedia (HT '08). ACM, New York, NY, USA, 55-60.

[6] Finanza Online Community. http://www.finanzaonline.it

[7] Gruhl D., R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2005. The predictive power of online chatter. In Proceedings of the eleventh ACM SIGKDD (KDD '05). ACM, New York, NY, USA, 78-87.

[8] Gu, B., P. Konana, A. Liu, B. Rajagopalan and J. Ghosh, 2007. Predictive value of stock message board sentiments. Working Paper, University of Texas at Austin.

[9] Spiegel U., T. Tavor, J. Templeman, 2010. "The effects of rumours on financial market efficiency," Applied Economics Letters, Taylor and Francis Journals, vol. 17(15), 1461-1464.

[10] Zhang, Y. and P.E. Swanson, 2009. Are day traders bias free?-Evidence from internet stock message boards. J. Econ. Finance. DOI: 10.1007/s12197008-90